AMENDMENTS TO THE CLAIMS:

1. (Currently amended) A method of improving at least one of speed and efficiency when executing a linear algebra subroutine on a computer having a memory hierarchical structure including at least one cache, said method comprising:

determining, for a level 3 matrix multiplication processing, which matrix will have data for a submatrix block residing in a lower level cache of said computer and which two matrices will have data for submatrix blocks residing in at least one higher level cache or a memory; and

streaming data from said selected two matrices ~~involved in processing~~ for executing ~~said linear algebra subroutine~~ said level 3 matrix multiplication processing ~~to a first matrix such that submatrix data of said two matrices residing in a higher level cache or in a memory is streamed to submatrix data of said first matrix residing in said cache,~~

~~said streaming providing data from said higher level to said data in said cache as required for said processing for executing correctly the linear algebra subroutine~~.

2. (Currently amended) The method of claim 1, wherein said ~~at least one~~ lower level cache comprises an L1 cache and said higher level cache comprises an L2 cache.

3. (Currently amended) The method of claim 1, ~~further comprising:~~ wherein ~~selecting~~ said determining said matrix ~~of the three~~ to be stored in said lower level cache ~~by~~ comprises ~~examining sizes and shapes of said three matrices~~ determining which of the three matrices has a smallest size.

4-5. (Canceled)

6. (Previously presented) The method of claim 2, wherein data for said second matrix and said third matrix streams into said L1 cache from said L2 cache such that said data from said second matrix and said third matrix streams in a vector format into said L1 cache.

7. (Currently amended) The method of claim 1, wherein said linear algebra subroutine comprises a substitute of a subroutine from ~~a~~ LAPACK (Linear Algebra PACKage).

8. (Currently amended) The method of claim 7, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

9. (Currently amended) An apparatus, comprising:

a memory system to store matrix data for a level 3 matrix multiplication processing ~~in a linear algebra program~~ using data from a first matrix, a second matrix, and a third matrix, said memory system including at least one cache; and

a processor to perform ~~a linear algebra operation~~ said level 3 matrix multiplication processing, wherein data from one of said first matrix, said second matrix, and said third matrix is stored as a submatrix block resident in ~~said~~ a lower level cache in a matrix format and data from a remaining two matrices is stored as submatrix blocks in said memory system at a level in said memory system higher ~~that~~ than said lower level cache,

said processor preliminarily selecting which matrix will have said submatrix block stored in said lower level cache and which said two matrices will have submatrix blocks stored in said higher level,

said data from said ~~remaining~~ selected two matrices being streamed through said lower level cache into said processor as required by said level 3 matrix multiplication processing ~~by streaming data from two of three matrices involved in processing said linear algebra subroutine to a first matrix such that submatrix data of said two matrices residing in a higher level cache or in a memory is streamed to submatrix data of said first matrix residing in said cache~~.

10. (Currently amended) The apparatus of claim 9, ~~further comprising:~~

~~a selector to determine~~ wherein said processor selects a ~~size of each of~~ smallest of said first, second, and third matrices ~~involved in a~~ to be said matrix ~~multiplication process and to select one of said matrices to reside~~ to have data residing in said first level cache~~, as based on having determined said sizes;~~

~~a loader to load data for the selected matrix into said cache ; and~~

~~a selector to select a matrix subroutine, from a plurality of~~ alternative matrix ~~subroutines, to perform said matrix multiplication process, each matrix subroutine in said plurality capable of executing said matrix multiplication process using one matrix operand as being cache resident and a remaining two matrix operands as streaming through said cache from said higher level cache or memory,~~

~~said selected matrix subroutine having a format consistent with which said matrix is selected to reside in said cache~~.

11. (Currently amended) The apparatus of claim 9, wherein said ~~linear algebra program~~ level 3 matrix multiplication comprises one or more subroutines substitute to ~~a~~ subroutine~~s~~ from a LAPACK (Linear Algebra PACKage).

12. (Currently amended) The apparatus of claim 11, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

13. (Canceled)

14. (Currently amended) A ~~signal-bearing~~ machine-readable storage medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of improving at least one of a speed and an efficiency of executing a ~~linear algebra subroutine~~ level 3 matrix multiplication processing on a computer having at least one lower level cache and one or more higher level caches or other higher level memory devices, said method comprising:

selecting which matrix will have submatrix block of data residing in said lower level cache and which two of three matrices will have submatrix blocks of data residing in at least one said higher level cache or memory; and

streaming data ~~for up to three~~ from said two selected matrices for said level 3 matrix multiplication processing ~~involved in processing said linear algebra subroutine such that data is processed using data for a first matrix stored in said cache as serving a matrix role in said linear algebra subroutine and being cache resident and data from a second matrix and a third matrix is stored at a higher level than said cache, said streaming providing data from~~

~~said higher level in a manner as said data is required for said processing by streaming data from said second and third matrices to submatrix data of said first matrix residing in said cache~~.

15. (Currently amended) The ~~signal bearing~~ machine-readable storage medium of claim 14, ~~said method further comprising:~~ wherein a smallest of said three matrices is selected as ~~determining a size of each of matrices involved in a matrix multiplication process;~~

~~selecting one of said matrices~~ said matrix to have data to reside in ~~one of~~ said ~~at least one~~ lower level cache~~, based on having determined said sizes;~~

~~arranging elements of elements of said matrices for said streaming; and~~

~~selecting a matrix subroutine from a plurality of subroutines by determining which said matrix subroutine can perform said matrix multiplication consistent with which matrix is selected to reside in said one of said at least one cache.~~

16. (Currently amended) The ~~signal bearing~~ machine-readable storage medium of claim 14, wherein said level 3 matrix ~~subroutine~~ multiplication processing comprises a substitute of a subroutine from LAPACK (Linear Algebra PACKage).

17. (Currently amended) The ~~signal bearing~~ machine-readable storage medium of claim 16, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

18. (Currently amended) The ~~signal bearing~~ machine-readable storage medium of claim 14, wherein said lower level cache comprises an L1 cache and data for said second

matrix and said third matrix streams from said higher level such that said data from one of said second matrix and said third matrix streams in a vector format ~~into~~ through said L1 cache.

19. (Currently amended) A method of providing a service involving at least one of solving and applying a scientific/engineering problem, said method comprising at least one of:

using a linear algebra software package that performs ~~one or more~~ a level 3 matrix multiplication processing ~~operations~~, said ~~method~~ software package comprising:

examining a size of each of three matrices involved in said level 3 matrix multiplication processing to select a smallest of said three matrices to have a block of submatrix data residing in an L1 cache and two remaining matrices to have data streamed from a higher level of memory; and

executing said level 3 matrix multiplication processing by said streaming of data ~~for~~ from said two selected matrices ~~involved in processing said linear algebra subroutines such that data is processed using data for a first matrix stored in a cache as a matrix format and data from a second matrix and a third matrix is stored in a memory device at a higher level than said cache, said streaming providing data from said higher level in a manner as said data is required for said processing by streaming data from two of three matrices involved in processing said linear algebra subroutine to a first matrix such that submatrix data of said two matrices residing in a higher level cache or in a memory is streamed to submatrix data of said first matrix residing in said cache~~;

providing a consultation for solving a scientific/engineering problem using said linear algebra software package;

transmitting a result of said linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing said result, and a printed version representing said result; and

receiving a result of said linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing said result, and a printed version representing said result.

20. (Canceled)

21. (New) The method of claim 1, said computer having M levels of caches and a main memory, said method further comprising:

selecting, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.

22. (New) The machine-readable storage medium of claim 14, said memory system including M levels of caches and a main memory, wherein said processor initially selects, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.

23. (New) The method of claim 1, wherein said computer comprises M levels of caches and a main memory, said method further comprising:

selecting, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.